Part 1

1.1. In our term-document matrix, the rows are word vectors of D dimensions. Do you think that's enough to represent the meaning of words? Why or why not?

I'd say no, probably not. Seeing as the vector only uses D features, if two words appear in identical sets of documents they'll have identical vectors (even if they're not co-occurring/showing up together).

1.2. Provide the top 10 associated words and cosine similarities (the output from `rank_words`) with `juliet` and at least 2 other target words of your choice for both term-document and term-context vector spaces.

word	matrix type	10 most similar words
juliet	td_matrix	1: juliet; 1.0 2: romeo; 0.9899494936611666 3: capulet; 0.9899494936611665 4: pump; 0.9899494936611665 5: laura; 0.9899494936611665 6: pitcher; 0.9899494936611665 7: behoveful; 0.9899494936611665 8: hurdle; 0.9899494936611665 9: capulets; 0.9899494936611665 10: petrucio; 0.9899494936611665
juliet	tc_matrix	1: juliet; 1.0 2: lucius; 0.787796461449417 3: gloucester; 0.7818061482037952 4: servants; 0.7717159769405332 5: warwick; 0.7677308387215129 6: nurse; 0.7590373891930818 7: paris; 0.7531720811811453 8: antonio; 0.7527362684737069 9: buckingham; 0.7489883918644319 10: brutus; 0.7485064965476168
yorick	td_matrix	1: flints; 1 2: umbrage; 1 3: uncharge; 1

		4: hinges; 1 5: betoken; 1 6: mutine; 1 7: historical; 1 8: rims; 1 9: sanctuarize; 1 10: flushing; 1
yorick	tc_matrix	1: yorick; 1.0 2: jester; 0.4654746681256313 3: soaks; 0.4364357804719847 4: damnably; 0.4330127018922194 5: alas; 0.4260903042986076 6: palace; 0.40903816554064876 7: wagered; 0.408248290463863 8: before; 0.4052921450216983 9: fidele; 0.4032795663087215 10: missed; 0.4003203845127178
wherefore	td_matrix	1: wherefore; 1.0 2: from; 0.9072035303013347 3: when; 0.9052750877016803 4: till; 0.9034815206620623 5: that; 0.89740313468159 6: thee; 0.8972149429417087 7: me; 0.8942687595870263 8: heart; 0.8913691996666683 9: where; 0.8912597111536651 10: patient; 0.8911772508175841
wherefore	tc_matrix	**1: wherefore; 1.0** **2: why; 0.8763188349501798 <-Interesting! A synonym!** 3: what; 0.8180559008889612 4: know; 0.8010787949209369 5: do; 0.8004844832462756 6: say; 0.7997776913925448 7: sir; 0.7974690511191451 8: how; 0.793213765495853 9: if; 0.7899890528436313 10: not; 0.7859003335211794

1.3. Just considering the term-document and term-context matrices without any tf-idf or PPMI weighting, which do you think produces similar words that make more sense than others? Why do you think that is the case? Back up your conclusions by referring to the top associated term lists you provided.

Definitely the term-context matrix (TCM). This was actually unreasonably fun, and quite interesting: I managed to generate a block of junk "rare words" with the Yorick TDM (a common pitfall), snagged Yorick's occupation in his TCM (jester), and got "wherefore"'s modern synonym in its TCM (why). The term-document matrices capture words that appear mostly in the same play, even if (contextually) they might not ever inhabit the same scene. A lot of these are gibberish. From the TCM's however I can identify all sorts of local cooccurrence patterns as mentioned above as these vectors pay heed to the terms that show up in similar contexts. This was a great demonstration, even with me picking my favorite character and a throwaway word at random.

1.4. Explain any decisions you made in implementing your functions, such as whether you allowed a target word to co-occur with itself as a context word, and which window size you chose for the term-context matrix. How might any decisions you make impact our results now?

I chose to throw in a little check to ensure that a word would not be counted as its own context, but it was mostly because I feel this wouldn't really *teach* us anything about a word's distributional properties. Co-occurring with itself is not too revelatory, I suppose? The window size is set to four when called and I did not adjust this - so, four words are considered to the left and right of a target.

Let's see: Narrowing the window would only capture the most immediate, local affiliations, and I'd wager the vectors would be sparser since there's less content to even have a chance at appearing in the window. Widening it probably makes things more dense and grab a few more associations that are less "smooshed up" against our terms of interest. Obviously, it would be more computationally intensive though.

Allowing a word to co-occur with itself I'm… not entirely sure I ever see being of interest or relevance to us? I'm not sure what's to be gained.

1.5. Provide the top 10 associated words (the output from `rank_words`) with `juliet` and at least 2 other target words of your choice for tf-idf-weighted term-document matrices and PPMI-weighted term-context matrices.

word	matrix type	10 most similar words

juliet	tf-idf	1: juliet; 1.0 2: mercutio; 0.9870437627628763 3: tybalt; 0.9870437627628763 4: pump; 0.9870437627628762 5: laura; 0.9870437627628762 6: pitcher; 0.9870437627628762 7: behoveful; 0.9870437627628762 8: hurdle; 0.9870437627628762 9: petrucio; 0.9870437627628762 10: heartless; 0.9870437627628762
juliet	ppmi	1: juliet; 1.0 2: capulet; 0.19196140184673016 3: vauntingly; 0.14946850506091336 4: barnardine; 0.140968929425194 5: provost; 0.13706838552940293 6: tybalt; 0.13685330053902423 7: montague; 0.13219117094886068 8: mercutio; 0.12983590962299896 9: romeo; 0.12548822860469921 10: stricken; 0.12338733674236768
yorick	tf-idf	1: flints; 1 2: umbrage; 1 3: uncharge; 1 4: hinges; 1 5: betoken; 1 6: mutine; 1 7: rims; 1 8: sanctuarize; 1 9: flushing; 1 10: winnowed; 1
yorick	ppmi	1: yorick; 1.0 2: jester; 0.2931973133306113 3: beholder; 0.2661894873996479 4: unbow; 0.25375008380965114 5: paunch; 0.23145776252289174 6: soaks; 0.22209369561098236 7: job; 0.22081545318690388 8: clothair; 0.2160360207490042 9: lain; 0.2026593265631963 10: undeserver; 0.1973325309240035
wherefore	tf-idf	1: wherefore; 1.0

		2: till; 0.9058695456575911 3: me; 0.8888135004720911 4: left; 0.8887050719509209 5: get; 0.8863978670144789 6: from; 0.8860713284035129 7: bear; 0.8823219288488422 8: break; 0.8823177474950755 9: patient; 0.8806590431634013 10: when; 0.8792904120491173
wherefore	ppmi	1: wherefore; 1.0 2: unthrifts; 0.16003405523376546 3: upstart; 0.142629079413497 4: incidency; 0.11288464430807021 5: caused; 0.10292675388589345 6: det; 0.10005558348612331 7: cometh; 0.09741441009288831 8: unpossessing; 0.0949833935863924 9: pander; 0.0938470369006481 10: afield; 0.09327481761132317

1.6. Compare the ranked word similarities between weighting with tf-idf and using the unweighted term-document matrix. Which do you think produces similar words that make more sense? Back up your conclusions with specific examples.

I would argue that TF-IDF seems to have done a better job. For “juliet”, we see “mercutio” and “tybalt” show up, bringing words that I would argue are more similar to the top. I’m not entirely sure what to think of “romeo” getting booted? For “yorick” the lists are basically the same - weighted or unweighted, I’m guessing it’s just a pile of single-occurrence words. It’s subtle with “wherefore”, but I suppose it seems like some common words have gotten shifted down a bit? “from” and “when” for example. Overall, yeah, I’m sticking with my initial thought: I find the weighting helpful and slightly more informative.

1.7. Compare the ranked word similarities between weighting with PPMI and using an unweighted term-context matrix. Which do you think produces similar words that make more sense? Back up your conclusions with specific examples.

Okay, this one is far less clear to me. It might be my favorite set: the PPMI weighting for “juliet” now shows a ton of characters related to her from the same play, whereas the plain old unweighted matrix just shows other “royal” characters, or folks with gravitas, from

other plays. The “yorick” matrices are a little less impressive - I’m sad to see “alas” vanish from the list, but it makes sense (a casual glance shows “alas” as appearing 10 times throughout Hamlet, and only once with regards to Yorick). It might be my fault for picking unhelpful, rare words, but “wherefore” also seems to decrease in quality.

To give PPMI the benefit of the doubt, I quickly reran things with “king” and “son”, which ought to be more illustrative:

```
The 10 most similar words to "king" using cosine-similarity on
term-context frequency matrix are:
1: king; 1.0
2: of; 0.9421805460074012
3: people; 0.9418049709002397
4: queen; 0.9382116211211731
5: french; 0.9380569448501763
6: dauphin; 0.9336528679071483
7: prince; 0.9322487017125826
8: devil; 0.9311691308915067
9: world; 0.9260394453734951
10: next; 0.9218125810521635

The 10 most similar words to "king" using cosine-similarity on PPMI matrix
are:
1: king; 1.0
2: henry; 0.22250468259421685
3: richard; 0.1583104238961972
4: lewis; 0.15581512129442354
5: queen; 0.14463713732826966
6: edward; 0.14262213731022266
7: enter; 0.14115437493089034
8: flourish; 0.13320237196470264
9: attendants; 0.1328494113582298
10: the; 0.13207923530968657

The 10 most similar words to "son" using cosine-similarity on term-context
frequency matrix are:
1: son; 1.0
2: father; 0.9449905348184818
3: daughter; 0.9420415342471975
4: mother; 0.9342339953878724
5: brother; 0.933394988060318
6: wife; 0.9261941320702185
7: life; 0.9080280676321102
8: mind; 0.9052429403765827
9: name; 0.8949589905482251
10: heart; 0.8927112751765767

The 10 most similar words to "son" using cosine-similarity on PPMI matrix
are:
1: son; 1.0
2: retourne; 0.15816716171290934
3: impudique; 0.15648919019036422
4: propre; 0.14666858216883005
```

```
5: contre; 0.1466131654355779
6: jurement; 0.14301667019533082
7: eldest; 0.13846300055871807
8: vomissement; 0.13301385797906806
9: encore; 0.13244117016325296
10: father; 0.1188390362940861
```

Okay! Well, maybe that didn't help. "king" sees several similar titles and roles in the TCM, which makes sense for words that share a context. Then for the PPMI we see specific names of kings. "son" we likewise see roles and titles that are similar, and then… it seems like its role as a possessive term in French is yielding a ton of random Frenchiness.

I'm going to have to go ahead and say, based on what I'm seeing, that while PPMI makes a good showing on occasion, it is genuinely the base TCM that impresses me more, making it my favorite of the four matrices.

1.8. Overall, do some approaches appear to work better than others, i.e produce better synonyms? Do any interesting patterns emerge? Discuss and point to specific examples.

Yes, some approaches definitely seem to work better in certain situations. No single approach seems to be the all-around winner for every use-case. Low-frequency terms get strange neighbors in the TCM and TDM, super common terms get unhelpfully inundated in the PPMI, and each of the term-document vectors struggle with single-play vocabulary.

I tried to call out interesting patterns as they came up, but overall "vibes" captured definitely include: "juliet" capturing general ritzy, Shakespearean names with the raw TCM but snagging specific family members and acquaintances with PPMI. "wherefore" in the same two matrices lost some value however - with TCM capturing "why" but PPMI probably booting it out for being overly frequent. "yorick" in the TDMs shows their weaknesses with rare words, but snagged "jester" in the TCM. "son" and "king" had the results I'd just shared, but the point is made: Each technique did very well in specific circumstances. Lots of interesting patterns of behavior all around.

Part 2

2.1 Provide the top 10 associated context words (by PPMI, in the SNLI corpus) for at least 4 identity labels of your choice. Choose identity labels that are related, such as multiple terms for gender, multiple terms for race/ethnicity, or other relations.

Identity Label	Top 10 PPMI
mexican	1: mexican; 1.0 2: tacos; 0.12233019320650784 3: restaurant; 0.12193243825932809 4: food; 0.11998361007723757 5: served; 0.11822345903206766 6: taco; 0.11703614794005635 7: traditional; 0.1131814685605641 8: seafood; 0.11278706877318012 9: festive; 0.11252120143406086 10: buffet; 0.11204410097418127
polish	1: polish; 1.0 2: nail; 0.4728386144712582 3: fingernail; 0.2344121663487907 4: wrinkles; 0.21091557758602997 5: fumes; 0.20592741666473025 6: toenails; 0.1927545259893635 7: fingernails; 0.18949425732531422 8: regrets; 0.188041341659272 9: polishing; 0.17890355009335857 10: nails; 0.1641960261896931
japanese	1: japanese; 1.0 2: chinese; 0.16391912798089725 3: traditional; 0.1634312938620086 4: kimonos; 0.13190339199650403 5: geisha; 0.1295985659175265 6: samurai; 0.12828566231425143 7: asian; 0.12260507958226674 8: warrior; 0.12225500286655033 9: colorful; 0.11414821670397746 10: costumes; 0.11396809330155078
african	1: african; 1.0

	2: american; 0.365434210604374 3: native; 0.16929623313362985 4: americans; 0.16021720104544146 5: tribe; 0.14863819732175032 6: asian; 0.14617216460886495 7: tribal; 0.14260155660543805 8: classroom; 0.13752359880899434 9: traditional; 0.12985926596252795 10: children; 0.1247273878973929

2.2. Do you see any associations learned by this bag-of-words model on the SNLI corpus that be representational harms, such as negative social stereotypes? Compare the top PPMI words for certain identity terms with other related ones (such as men compared with women). Discuss and provide selected results. If you don't find any representational harms (that's okay), provide examples of what you examined and how you interpreted those associations. If you do find problematic associations, specify how they could be harmful.

Oh, I absolutely see all sorts of connections that could be deemed harmful. I opted for a mix of cultural and national identities:

1. Mexicans are reduced entirely to terms surrounding celebrations and food, which is not the most obviously harmful but does highlight the extractive treatment of their people and culture by a lot of Western nations (your food is fine, but your people…?).
2. Africans have a ton of affiliations with "tribes" and "tribalism", infantilizing concepts with a major primitive bent to them that are explicitly harmful (while not words that I included, "indian" was another interesting one to spot check - a culture being superseded by Western depictions of backwards, uncivilized populaces with a simultaneous disregard for the geographical truth of the term *and* the modern context of their involvement).
3. Japanese is linked with a number of romanticized concepts from the West, like geisha and samurai, perpetuating beliefs about submissive women and emotionally stoic men that have been harmful for many.
4. Polish I included for a bit of levity, but it is interesting and noteworthy that the hundreds of years of Polish identity are overshadowed and overwritten by talk of fingernails, toenails, and regrets.

There are many forms that harm can take, and I think this exercise does a good job highlighting just how much just how easy it is for sensible math and genuinely effective text parsing in a literary context to suddenly inherit the biases and nonsense of our real

world when applied to a more nebulous, real-world dataset. Anyone who argues there is not a place for ethical discussions in data science is truly misguided.

2.3. For at least 4 pairs of identity terms and highly associated words, provide the document contexts in the SNLI dataset that contribute to this association. Provide actual sentences from the SNLI corpus where either:

- **an identity term occurs together with the associated word you found (1st-order similarity) or**
- **an identity term occurs separately from the associated word, but occurs with similar context words (2nd-order similarity). You'll want to compare values of dimensions in the vectors for both words in this case.**

Discuss any findings, particularly related to any associations you found to represent harmful stereotypes.

1. **"mexican" & "tacos"** - Interestingly, it seems that 2nd-order similarities dictate this connection by and large. Sentence "c" sees a direction 1st-order association, but by and large talk of tacos is predictably almost always in a dining context without any mention of Mexicans. Talk of Mexicans on the other hand is very often in a context of poverty, crime, or - as we see below - food.
    a. A man is sweeping the street in front of a restaurant called Tacos La Palapa.
    b. The soccer match ends with superman eating tacos
    c. People ordering tacos at a Mexican take-out spot
    d. some tacos ride a bike.
    e. A man is waiting in line to get the famous panda meat tacos.
    f. The food is mexican.
    g. The table is set with many types of Mexican dishes.
    h. A waiter at a Mexican restaurant.
    i. People laughing while eating Mexican food.
    j. The restaurant is serving Mexican food.
    k. The restuarant serves Mexican food.

2. **"african" and "tribe"** - Lots of associations here, both 1st-order and 2nd-order. For the latter we see so many terms: barefoot, hut, villagers, poor, garb, tribals. I particularly enjoy the sentence where African clothing is "garb" while Western clothing is "wear" (we really don't ever see references to Caucasian garb). I had to stop eventually so as to not overpopulate this, but there are plenty of 1st-order associations. Any group of black people in a rural or underdeveloped setting are immediately a "tribe".
    a. Some barefoot African boys are wearing containers.
    b. three young african children in a poor part of africa one riding a bike.

c. Three black people dressed in african garb watching something.
d. An older woman dressed in the garments of an African tribe has her right hand in a cracked brown pot and holds her left hand over another brown pot.
e. A group of African villagers are carrying an animal using their hands to tie it up.
f. Women in African garb sit outside while one in Western wear walks past.
g. African tribals carrying guns.
h. Two African children standing by a hut with their faces painted.
i. Africans are gathering around huts talking.
j. People from an African tribe are gathered outside.

3. **“japanese” and “geisha”** - Okay, first off, I apologize for the unrelated sentences, but I really enjoy the number of captions that are calling out “Japanese or Chinese” in contexts where I’m certain a more nuanced analysis would strictly separate the two for cleaner and better specified data.
Anyway, I believe we again see a mixture here. Plenty of 1st-order references to Japanese geisha, but also a strong 2nd-order set of connections between kimonos, dresses, makeup, Japanese women, and geisha as a conglomeration. While traditionally more theatrical and artistic, there was a strong shift towards an association with sex work and prostitution that is harmful to so strongly connect to the culture at large.
    a. A Japanese woman in a colorful outfit is walking with a man dressed in black past a wooden building.
    b. Two women in Japanese dress walk down the street.
    c. Two women in elaborate Japanese costumes.
    d. Two Japanese ladies with colorful kimonos on them.
    e. Two woman dressed as geishas are riding in a cart pulled by a man.
    f. A man pulls a coach with two Geishas on board.
    g. A geisha stands outside of a store peering to her side.
    h. A young woman with blond-hair is holding a white statuette of a Japanese geisha woman.
    i. Two geisha happy geishas wearing pink kimonos.
    j. The 3 Asian girls are pretending to dress up like Geisha girls.
    k. A woman dressed up as a geisha.
    l. geisha loves makeup
    m. Two young geisha wearing bright red kimonos pose for tourists.
    n. Two Japanese people dressed in kimonos are geishas
    o. "The geisha had the good sense to laugh, for the client was fearsome and easily angered."

4. **“polish” and “fingernails”** - I’ll go ahead and include the singular reference to Poland I could find below. Universally, everything else is dealing with the act of polishing nails and shoes. So… what do I call this? I suppose these are all 1st-order

with regards to the act of polishing, but also 1st-order displacing/erasing associations with anything Polish? This is just linguistic ambiguity at play, but simultaneously a dataset that is lacking in anything truly reflective of Poland in its scarcity. On the bright side, I figured out why “regrets” features so heavily - there are multiple photos of a woman showing off her regrets tattoo while sporting nail polish.

a. The restaurant supporst the Polish president.
b. Woman polishing her nails.
c. A man likes to collect nail polish posters
d. "Hi, Which are you using L'Oreal nail polish on your fingers.?"
e. "A girl with blue nail polish, is in a workshop or garage among a variety of tools and objects and is pouring a clear liquid from a large, clear pitcher into a green, two-liter soda bottle."
f. "A woman wearing blue nail polish is showing her ""no regrets"" tattoo to the camera."

This assignment was very enjoyable and demonstrative of both the strengths and weaknesses of these rudimentary techniques. The tools to meaningfully extract information from a large corpus are readily available, but our stereotypes (intentional or otherwise) are at risk of shining through. Thank you again!